社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2008－MPS－72 (20)
2008－BIO－15 (20)
2008／12／17

# 多目的遺伝的アルゴリズムによるSVM学習データ選択手法

廣安 知之[†], 西岡 雅史[††], 三木 光範[‡], 横内 久猛[†]

[†] 同志社大学生命医科学部 [††] 同志社大学大学院工学研究科 [‡] 同志社大学理工学部

サポートベクターマシン（SVM）の学習では，どのデータを学習データとして扱うかの選択が重要となる．これは，多くの学習データを利用する際に過学習の問題が存在するからである．ユーザは，どれだけの学習データを利用するのかを決定し，与えられたデータセットから学習データを選択しなければならない．本研究では，このSVM学習データの選択を多目的最適化問題としてとらえ，多目的遺伝的アルゴリズム（多目的GA）を適用することによって最適化する．このとき，問題の特徴を把握するためには幅広いパレート解集合を導出する必要がある．したがって，我々の提案する多目的GAのための探索戦略を，学習データの選択に適用することが有効であると考えられる．提案する探索戦略をSVM学習データ選択問題へと適用させた結果，従来の多目的GA手法に比べて幅広い解集合を導出可能であることが確認できた．また，探索戦略を用いることにより，SVMの学習における学習データの分類性能と汎化能力のトレードオフ関係をより正確に把握することができた．

# SVM Training Data Selection Using Multi-Objective Genetic Algorithm

Tomoyuki HIROYASU[†] Masashi NISHIOKA[††], Mitsunori MIKI [‡], Hisatake Yokouchi[†]

[†] Faculty of Life and Medical Sciences, Doshisha University
[††] Graduate School of Engineering, Doshisha University

[‡] Faculty of Science and Engineering, Doshisha University

When training Support Vector Machine (SVM), selection of a training data set becomes an important issue, since the problem of overfitting exists with a large number of training data. A user must decide how much training data to use in the training, and then select the data to be used from a given data set. We considered to handle this SVM training data selection as a multi-objective optimization problem and applied our proposed MOGA search strategy to it. It is essential for a broad set of Pareto solutions to be obtained for the purpose of understanding the characteristics of the problem, and we considered the proposed search strategy to be suitable. The proposed search strategy was adapted to the SVM training data selection problem, and the results of the experiment indicated that broader solutions can be obtained by the search strategy compared to conventional MOGAs. Moreover, better understanding of the tradeoff relationship between the classification performance of the training data and the generalization performance became possible with the proposed search strategy.

## 1 はじめに

サポートベクターマシン（SVM）[1]は，V. Vapnikなどによって提案されたパターン識別手法である．SVMにおける学習の目標は，未知のデータを正確に分類する汎化能力の向上である．そのためには効果的な学習が必要であるが，このとき考慮しなければならないのが過学習である．過学習とは，学習データに対する分類性能を必要以上に向上させることにより，決定関数が複雑になり，汎化能力が低下する現象である．したがって，汎化能力の高いSVMを実現するためには，どのような場合に過学習が起こっているのかを把握し，対象となるパターン認識問題の特徴を理解する必要がある．

そこで，本稿ではSVMの学習に複数の目的の下で最適化を行う，多目的最適化手法を適用させることによって，SVMの学習の特徴を把握する．具体的には，SVMの学習に用いる学習データの選択を対象問題とし，学習データに対する分類性能と汎化能力を最適化する2目的最適化問題として扱う．学習データの最適化には，多目的遺伝的アルゴリズム（多目的GA）を用いる．このとき，問題

の特徴を正確に把握するためには，各目的における最適解を精度良く導出する必要がある．そのため，一般的な多目的 GA 手法よりも幅広い解集合を導出することができる探索戦略を提案し，SVM 学習データ選択問題へと適用させる．

## 2 多目的最適化

### 2.1 多目的最適化問題

複数の評価基準が存在し，これらの評価基準を同時に考慮しながら最適化を行う問題を，多目的最適化問題という．一般に多目的最適化問題では，複数の目的関数同士が互いにトレードオフの関係にある場合が多いため，全ての目的関数 $f_i(\boldsymbol{x})$ を同時に最適化することはできない．そこで，他のどのような解にも劣らない，パレート最適解を求めることが目標となる．

一般にパレート最適解は複数存在することが多く，多目的最適化問題ではパレート最適フロント全域を覆う解集合を導出することが目標となる．したがって，導出される解集合は精度，均一性，幅広さの全ての要素について優れていることが望ましい．精度とは，得られた解集合がパレート最適フロントにどれだけ近いかであり，均一性とは解集合が特定の領域に偏ることなく，均一に分布しているかどうかである．また，幅広さとは解集合がどれだけ広い領域に分布しているかであり，これはパレートフロントの端に位置する解，すなわち各目的における最適解によって決定される．

### 2.2 多目的遺伝的アルゴリズム

多目的最適化の分野では，様々な進化的計算手法が適用されているが，特に遺伝的アルゴリズム（GA）を多目的最適化に適用した多目的 GA は最も多く研究されている．現在，代表的な多目的 GA 手法として，Deb らの NSGA-II や Zitzler らの SPEA2 などがあり，良好な性能を示すことが知られている．

## 3 精度と幅広さを考慮した多目的遺伝的アルゴリズムのための探索戦略

SVM 学習データ選択問題において，学習データに対する分類性能と汎化能力の間に存在する，トレードオフ関係を正確に把握するためには，精度と幅広さに優れた解集合が必要である．しかしながら，一般的な多目的 GA 手法では，得られた解集合の幅広さを向上させるためのメカニズムが組み込まれていないことが多い．そこで，本稿では解集合の精度と幅広さに注目した，多目的 GA のための探索戦略を提案する．

奥田らが提案した分散協力型スキーム [2] は，解集合の幅広さを向上させることができるが，一方で探索の精度が低下することがわかっている．このように，探索中に精度と幅広さを同時に向上させることは困難である．そこで，提案する探索戦略では図 1 に示すように，探索を 2 段階に分割することを考える．探索の 1 段階目では精度を向上させ，2 段階目では解集合の幅広さを向上させる．

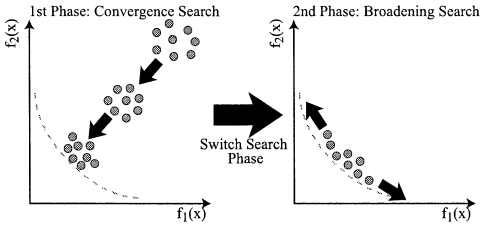

図 1: 提案する探索戦略の概念図

### 3.1 1 段階目：精度を重視した探索

1 段階目の精度を重視した探索では，意思決定者の選好情報を希求点として利用する．一般的な多目的 GA 手法では，解の優越関係によって探索が進行するのに対し，提案手法では優越関係と希求点からの距離情報を用いて探索を進める．提案手法は一般的な多目的 GA 手法を基にしており，希求点からの距離情報は探索母集団を選択するメイティング選択において，選択基準の一つとして利用する．アーカイブサイズ $N$ とするときの，メイティング選択手法を下記に示す．

**Step 1:** アーカイブ内の解を希求点からのユークリッド距離によって昇順にソートする．

**Step 2:** 希求点に近い上位 $\frac{N}{2}$ 個の解を探索母集団に加える．

**Step 3:** 残りの解をランクによるトーナメント選択によって選択する．同一なランクを持った解が複数存在する場合，希求点からのユークリッド距離が最も近い解を選択する．

### 3.2 2 段階目：幅広さを重視した探索

2 段階目の幅広さを重視した探索には，奥田らの提案した分散協力型スキーム [2] を用いる．分散協力型スキームでは，探索母集団を単一目的 GA で探索するサブ母集団（SOGA 個体群）と多目的 GA で探索するサブ母集団（MOGA 個体群）に分割する．$k$ 目的最適化問題の場合，探索母集団は 1 つの MOGA 個体群と $k$ 個の SOGA 個体群に分割され，合計で $k+1$ 個のサブ母集団が形成される．分散協力型スキームの概念図を図 2 に示す．

分散協力型スキームにおいて MOGA 個体群と SOGA 個体群は並列に探索し，各個体群の最良解

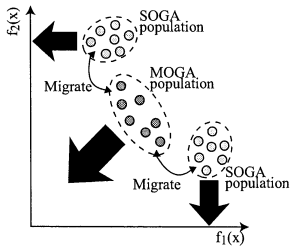

図 2: 分散協力型スキームの概念図

があらかじめ定められた世代数毎に交換される．なお，本稿では 25 世代毎に最良解の交換を行い，最良解の交換は MOGA 個体群と各 SOGA 個体群の間のみで行う．

### 3.3 探索の切り替え

提案する探索戦略では，1 段階目の探索から 2 段階目へと切り替えるタイミングが重要となる．精度を重視した探索から幅広さを重視した探索に切り替えるタイミングは，探索が十分に収束した時であることが望ましい．ここでは，以下に示す 2 通りの場合に探索が十分に収束したと判断する．

- パレート最適フロントへの収束が停滞している．
- 新たに生成された非劣解がパレート最適フロントへの収束に貢献していない．

したがって，本研究では探索の収束具合を表す 2 つの指標を用いて探索を切り替える．1 つ目の指標には Jaimes らの MRMOGA で用いられている指標を利用する．この指標では，探索中におけるアーカイブ内の非劣解が，どの程度の割合で次世代の解によって優越されるかを世代ごとに計測することで，解集合が収束しているかを判断する．

2 つ目の指標では，新たに生成された非劣解が優越するアーカイブ内の非劣解の平均数を扱う．この指標を用いることにより，MRMOGA の指標では考慮されていない新たな非劣解の数を考慮することができる．優越する解の平均数が低い場合，探索は多様性の向上へと移行していると考えられ，十分に収束していると判断できる．上記の 2 つの指標のいずれかの条件が満たされた場合に，提案する探索戦略では探索を切り替える．

## 4 SVM 学習データ選択問題への適用

### 4.1 サポートベクターマシン

サポートベクターマシン（SVM）は，V. Vapnik などによって提案された，パターン認識の分野において優れた性能を示すことが知られている手法である [1]．近年，カーネル関数の導入によって線形分離可能でないデータに対しても，優れた性能を示す非線形 SVM が多く研究されている．なお，本稿ではカーネル関数として式（1）に示す RBF カーネルを用いる．

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\gamma||\boldsymbol{x}_i - \boldsymbol{x}_j||^2\right), \gamma > 0 \quad (1)$$

### 4.2 SVM 学習データ選択問題

本論文では，3 節で述べた探索戦略を，SVM 学習データの選択へ適用する．一般に SVM の学習を行う場合，あらかじめ与えられたデータセットを学習データと検証データに分類する必要がある．学習データに対する SVM の分類性能は学習データの増加に伴い向上すると考えられるが，同時に過学習によって汎化能力が低下する可能性もある．そこで，SVM 学習データ選択問題における目的を，以下の 2 つとして定義する．

- 学習データに対する誤識別率（$f_1$）
- 検証データに対する誤識別率（$f_2$）

両目的は最小化するものとする．学習データに対する誤識別率を改善することは，汎化能力の低下をともなう過学習につながると予想され，この 2 つの目的間にはトレードオフ関係が存在すると考えられる．このとき，各目的における最適解を導出することが，SVM による学習の特徴を把握する上で重要となる．

### 4.3 実装方法

この問題は多目的ナップサック問題と同様の方法で実装されており，データセットに含まれる各データが学習データとして使われるかは，0/1 のビットで表される．ビットが 1 の場合には対応するデータを学習データとし，0 である場合には検証データとする．したがって，全てのデータに対応するビットを 0/1 で表現することにより，学習データの数とどのデータを学習データとするかを同時に決定することができる．

様々なデータセットを用いて対象問題を実装したが，本稿では特徴的な結果を示した，表 1 に示す Diabetes，Australian と Vehicle をデータセットとして用いる．対象とする SVM は $C$-SVM とし，RBF カーネルを用いる．なお，SVM のパラメータである $C$ と $\gamma$ は事前に交差検定によって求める．

表 1: 実験に用いるデータセット．$n$ はデータ数，$m$ は特徴の数．

| Data set | $n$ | $m$ | classes | $C$ | $\gamma$ |
|---|---|---|---|---|---|
| Diabetes | 768 | 8 | 2 | 32.00000 | 0.03125 |
| Australian | 690 | 14 | 2 | 0.03125 | 0.03125 |
| Vehicle | 846 | 18 | 4 | 128.00000 | 0.12500 |

問題における制約条件として，学習データの数は全体のデータ数の 1/2 以下とする．多目的 GA

手法のパラメータとして，母集団サイズ 120，最大世代数 250 とする．したがって，評価計算回数は同じである．他にも，交叉には 2 点交叉を用い，交叉率 1.0，突然変異率は 1/染色体長 とする．探索戦略で用いる分散協力型スキームで利用する DGA のパラメータは，サブ母集団サイズ 10，エリート数 1，トーナメントサイズ 4 とする．移住トポロジはランダムリングであり，移住率 0.5，移住間隔 5 世代とする．また，探索戦略で用いる希求点については，2 目的の最小化問題であるため，$(0,0)$ に設定する．

### 4.4 実験結果

図 3 と図 4 に，Diabetes，Vehicle における 50% 到達領域を示す．探索結果が示すように，2 つの目的間にはトレードオフ関係がいずれのデータセットにおいても見られた．

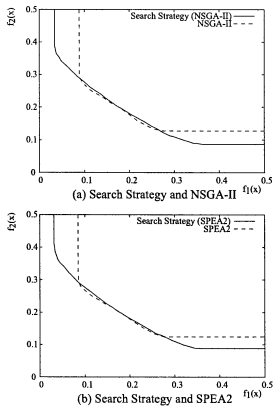


図 3: Diabetes における 50%到達領域（30 試行）

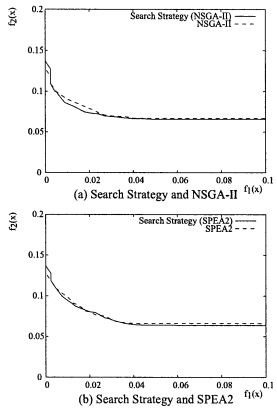


図 4: Vehicle における 50%到達領域（30 試行）

図 3 に示した Diabetes の探索結果では，探索戦略は NSGA-II や SPEA2 に比べて幅広い解集合を導出している．これは，意思決定者がパレートフロントの形状およびにトレードオフの度合いを把握するために有用な情報であるといえる．その一方で，図 4 に示した Vehicle の探索結果には，明確な違いは見られなかった．

学習データの数に注目してみると，Diabetes と Australian のデータセットについては，探索戦略と一般的な多目的 GA 手法の間で，データ数に違いが見られなかった．しかしながら，図 3 に示したように，Diabetes について実際に得られた解集合は，幅広さに大きな差がある．各目的における最適解が大きく異なることから，学習データに対する分類性能と検証データに対する分類性能は，データ数だけによって決まらないことがわかる．

一方で，Vehicle のデータセットでは学習データ数が大きく異なっていた．しかし，図 4 に示したように，実際に得られた解集合は同等である．つまり，学習データの選択方法によっては，同様な性能を示す SVM がより少ない，もしくは多い学習データ数によって得られることがわかる．この結果からも，学習データの選択が SVM の性能に大きく影響することが確認できる．また，選択した学習データに応じた SVM パラメータの最適化が必要であると考えられる．

## 5 終わりに

本稿では，SVM における学習データの選択を多目的最適化問題としてとらえ，多目的 GA を用いた探索戦略を適用させた．適用させた探索戦略では，探索を 2 段階に分割することによって解集合の精度と幅広さをそれぞれ向上させる．実験結果から，探索戦略を用いることによってより幅広い解集合を得られることが確認できた．これにより，問題の特徴を理解しやすくなり，どの程度のトレードオフが目的間に存在するかを把握することができる．また，解集合に含まれる解の学習データ数についても，探索戦略と一般的な多目的 GA 手法では違いが見られ，SVM の学習におけるパラメータの最適化の必要性を把握することができた．学習データによって最適な SVM パラメータは異なると考えられるため，今後の研究では学習データの最適化と同時に，SVM パラメータの最適化を行うことを検討する．

## 参考文献


1) Corinna Cortes and Vladimir Vapnik: Support-Vector Networks, *Machine Learning*, Vol. 20, No. 3, pp. 273–297 (1995).
2) Tamaki Okuda, Tomoyuki Hiroyasu, Mitsunori Miki and Shinya Watanabe: DCMOGA: Distributed Cooperation model of Multi-Objective Genetic Algorithm, *Proc. Advances in Nature-Inspired Computation: The PPSN VII Workshops*, pp. 25–26 (2002).